BAYES FACTOR BASED SPEAKER CLUSTERING FOR SPEAKER DIARIZATION

D. Wang, R. Vogt and S. Sridharan

Speech and Audio Research Laboratory,
Queensland University of Technology,
Brisbane, Australia
di.wang@student.qut.edu.au, r.vogt@qut.edu.au, s.sridharan@qut.edu.au

ABSTRACT

This paper proposes the use of the Bayes Factor to replace the Bayesian Information Criterion (BIC) as a criterion for speaker clustering within a speaker diarization system. The BIC is one of the most popular decision criteria used in speaker diarization systems today. However, it will be shown in this paper that the BIC is only an approximation to the Bayes factor of marginal likelihoods of the data given each hypothesis. This paper uses the Bayes factor directly as a decision criterion for speaker clustering, thus removing the error introduced by the BIC approximation. Results obtained on the 2002 Rich Transcription (RT-02) Evaluation dataset show an improved clustering performance, leading to a 14.7% relative improvement in the overall Diarization Error Rate (DER) compared to the baseline system.

1. INTRODUCTION

Speaker diarization systems have proven useful in areas such as speaker indexing and information retrieval as well as assisting in speech recognition applications. In information retrieval applications, a speaker diarization system allows automatic indexing of spoken audio documents, enabling the end user to browse the audio document by speaker. In speech recognition applications, speaker diarization can be used to localize the instances of a specific speaker to pool data for model adaptation, which in turn boosts transcription accuracies. Speaker diarization hence plays an important role in automatic transcription of broadcast news [1, 2].

Speaker clustering, the process of associating segments of speech produced by the same speaker, is commonly regarded as the most crucial step in the final stages of a speaker diarization system and is the focus of this paper. The use of the Bayesian Information Criterion (BIC) for speaker clustering was first proposed by the pioneering work of Chen *et al.* [3]. Chen proposed an agglomerative clustering algorithm using the BIC as a decision criterion, to determine whether two audio segments came from the same speaker. In [3], multivariate Gaussians are chosen to model the utterances, and the BIC is used as a model selection criterion to determine whether the two segments of interest are more appropriately represented by a single Gaussian or two separate Gaussians. BIC based clustering approaches have since received increasing acceptance in the speech technology community [4], and has become one of the most popular clustering strategies to date due to its robustness and threshold independence. Published speaker diarization systems that use the BIC for speaker clustering include the LIMSI broadcast news diarization system [1], which was the top participant in the most recent NIST Rich Transcription broadcast news evaluation, the RT-04F [5]. The baseline system used for comparison in this paper is based on this system.

However, within the statistical hypothesis testing framework, the BIC can be seen as only an approximation to the Bayes Factor of marginal likelihoods of the data given each hypothesis. The BIC approximation neglects the crucial prior term in the estimation of the marginal probability of the data given the model, thus foregoing the ability to incorporate prior beliefs about the expected distribution of parameter values. There is hence no guarantee that the marginal probability calculated using the BIC approximation would be close to the "true" value, calculated using a prior distribution that would be regarded as appropriate by an observer. This paper presents a speaker clustering technique based on the Bayes Factor itself, with the aim of improving clustering performance and consistency by removing the errors caused by the BIC approximation.

Section 2 gives an overview of the baseline broadcast news diarization system, which incorporates a BIC clustering stage. Section 3 presents the Bayes Factor of marginal likelihoods of the data given each hypothesis as a decision criterion for speaker clustering, and contrasts the BIC approximation to the actual marginal likelihood expression proposed in this paper. Section 4 presents the result obtained on the RT-02 Evaluation dataset and compares the result to the baseline system, and Section 5 draws some conclusions.

2. BASELINE SYSTEM OVERVIEW

The baseline system used for comparison in this paper is based on the **c-sid** configuration of the LIMSI broadcast news diarization system [1]. In the baseline system, the audio is first passed through a speech activity detection stage which separates the audio into speech and non-speech regions. Speaker segmentation is then performed to partition the speech regions into homogeneous speaker segments. This is followed by a Viterbi resegmentation stage which aims to refine the segment boundary locations. The set of speaker segments are then passed to the speaker clustering stages of the system, which aim to merge the segments containing the utterances produced by the same speaker.

Speaker clustering is performed in two separate stages, a BIC based initial clustering stage followed by a second clustering stage based on a speaker identification (SID) method. Both clustering stages use agglomerative clustering, where clustering is performed by iteratively merging the closest pair of clusters. Due to the lack of data in the initial clustering stage, where speaker segments are relatively short, a multivariate normal distribution is used to model the data, as opposed to a Gaussian Mixture Model (GMM). The initial clustering stage merges only the closest speaker segments and is terminated early, resulting in a set of underclustered nodes, which is passed into the second clustering stage that performs further clustering using more complex models. The performance of the initial clustering stage is hence crucial to the success of the overall diarization system, since correct clustering decisions made in this stage will generate pure, homogeneous clusters with sufficient data to be represented by more complex models in the subsequent clustering stage.

At the end of the initial clustering stage, the segment boundaries are refined once more via Viterbi resegmentation. The refined segments are then passed into the SID clustering stage, which completes the clustering process. At this stage, the initial clusters have considerably more data than the individual speaker segments passed into the BIC clustering stage. GMMs are therefore used to model the complex distribution of data in each speaker cluster. The SID clustering stage produces the final diarization output, consisting of a relative, show-internal set of speaker labels and their corresponding start and end times.

3. THE BAYES FACTOR AS A DECISION CRITERION FOR SPEAKER CLUSTERING

This section presents the Bayes Factor of marginal likelihoods of the data given each hypothesis as a decision criterion for speaker clustering. A derivation of the BIC is presented as an approximation to the marginal likelihood integral, and its shortfalls are outlined. This is followed by a derivation of an exact expression of the marginal probability integral to construct the Bayes factor for multivariate Gaussian models as a decision criterion for speaker clustering.

3.1. The Bayes Factor of Marginal Likelihoods

To derive an expression for the Bayes Factor as a decision criterion for model selection in the framework of hypothesis testing, let the null hypothesis, H_0, be that the two segments are more appropriately modelled by one multivariate Gaussian distribution (and hence should be clustered), and the alternative hypothesis, H_1, be that the two segments should be modelled by two separate Gaussian distributions (and hence should be kept separate). Let the data to be modelled be given by $\boldsymbol{X} = \{\boldsymbol{x}_i : i = 1, \cdots, N\}$. According to Bayesian decision theory, the criterion based on which the clustering should be made is given by

$$\frac{p(H_0|\boldsymbol{X})}{p(H_1|\boldsymbol{X})}\,. \tag{1}$$

Applying Bayes Theorem, the posterior probability of each hypothesis given the data can be written as

$$p(H|\boldsymbol{X}) = \frac{p(H)p(\boldsymbol{X}|H)}{p(\boldsymbol{X})}\,. \tag{2}$$

Since $p(\boldsymbol{X})$ is identical for both hypotheses, it will not affect the hypothesis testing and will cancel out under the decision criterion given in (1). Also, assuming equal prior probability for each hypothesis (ie. $p(H_0) = p(H_1) = \frac{1}{2}$), $p(H)$ will also cancel out. Under these assumptions, $p(H|\boldsymbol{X})$ is proportional to the likelihood of the data given each hypothesis, and the Bayes Factor, B, defined as a ratio of the likelihood of the data given the two competing hypotheses, can be written as

$$B = \frac{p(\boldsymbol{X}|H_0)}{p(\boldsymbol{X}|H_1)}\,. \tag{3}$$

Let the first speaker segment contain data $\boldsymbol{X}_1$ and the second contain $\boldsymbol{X}_2$. Let the single multivariate Gaussian distribution that supports H_0 be M_0, and the separate Gaussian distributions that support H_1 be M_1 and M_2 respectively. The Bayes Factor given in (3) can then be written as

$$B = \frac{p(\boldsymbol{X}|H_0)}{p(\boldsymbol{X}|H_1)} = \frac{p(\boldsymbol{X}_1, \boldsymbol{X}_2|M_0)}{p(\boldsymbol{X}_1|M_1)p(\boldsymbol{X}_2|M_2)}\,. \tag{4}$$

As evident from (4), the larger the Bayes Factor, the more evidence that the two segments are more appropriately modelled by one multivariate normal distribution and should be clustered, and vice versa.

To evaluate the Bayes Factor, one must first derive an expression for each of the terms on the right hand side of (4). Let $\boldsymbol{\lambda}$ be the parameter set of the model under consideration. The probability that the data conform to a model M, can be given by the marginal probability integral

$$p(\boldsymbol{X}|M) = \int p(\boldsymbol{X}|\boldsymbol{\lambda}, M)p(\boldsymbol{\lambda}|M)\, d\boldsymbol{\lambda}\,. \tag{5}$$

The marginal probability can be interpreted as the expected value of the likelihood of the data given the model. It is given by the likelihood of the data given each model parameter set, $p(\boldsymbol{X}|\boldsymbol{\lambda}, M)$, weighted by the associated prior probabilities of each particular set of model parameters, $p(\boldsymbol{\lambda}|M)$, and integrated over all possible values of the parameters.

3.2. The BIC approximation

The marginal probability integral shown in (5) is difficult to compute when M has a large number of parameters, for example when M is a high-order GMM, as it involves integrating over a large number of parameters, and difficulties arise in deciding which data belongs to which mixture component. The BIC was hence introduced as a relatively simple approximation to the Bayes Factor, by using the Laplace Approximation [6] to derive an expression for $\log p(\boldsymbol{X}|M)$. The Laplace Approximation aims to approximate a probability density function defined over a set of continuous variables by finding a Gaussian approximation centred on a mode of the distribution. Using the Laplace Approximation, (5) can be written as

$$\begin{aligned}\log p(\boldsymbol{X}|M) = \log & p(\boldsymbol{X}|\boldsymbol{\lambda}_{\text{MAP}}, M) + \log p(\boldsymbol{\lambda}_{\text{MAP}}|M) \\ & + \frac{k}{2}\log(2\pi) - \frac{k}{2}\log(N) - \frac{1}{2}\log|\boldsymbol{J}|\,,\end{aligned} \tag{6}$$

where $\boldsymbol{\lambda}_{\text{MAP}}$ is the value of $\boldsymbol{\lambda}$ at the mode of the posterior distribution (ie. $\log p(\boldsymbol{X}|\boldsymbol{\lambda}_{\text{MAP}}, M)$ is the maximum log likelihood of the data under model M), $\boldsymbol{J}$ is the expected information matrix for a single observation and N is the number of data points.

Ignoring the second, third and last term in (6) gives the BIC approximation for the marginal log likelihood of the data given model M [7],

$$\text{BIC} = \log p(\boldsymbol{X}|\boldsymbol{\lambda}_{\text{MAP}}, M) - \frac{k}{2}\log(N)\,. \tag{7}$$

As the above mathematical derivation shows, the BIC is only an approximation to the marginal probability integral. Given a constant number of model parameters, the third term of (6) is simply a constant offset and can hence be ignored without affecting the model selection. The effect of ignoring the last term is difficult to generalize, as it depends on the nature of the data as well as the parameterization. The biggest shortfall of the BIC approximation appears to be a result of ignoring the second term, which takes into account the prior beliefs about the expected distribution of parameter values. Ignoring the prior term effectively means that the ability to incoporate one's prior beliefs about the expected distribution of parameter values has been lost, and there is no guarantee that the value of the marginal likelihood calculated from the BIC approximation will be close to the "true" marginal likelihood calculated from a prior distribution that an observer would regard as appropriate for the data being modelled [7].

3.3. The Exact Value of the Marginal Probability Integral

Due to the shortfalls of the BIC approximation outlined in the above section, this paper proposes that it will be advantageous to compute the exact value of the marginal probability integral, shown in (5), to give the best possible estimate for $p(\boldsymbol{X}|M)$.

The exact values of the marginal likelihood of the data given the model can then be used to construct the Bayes Factor as a decision criterion for speaker clustering.

To evaluate the exact value of the integral in (5), one must first choose an appropriate distribution for $p(\boldsymbol{\lambda}|M)$ to reflect one's prior beliefs about the expected distribution of parameter values. Following from common practice in speaker recognition and for simplicity, this paper will consider only the means of the distributions as the parameters in evaluating the marginal likelihood integral. The prior chosen here is hence the prior on the mean. The variances will be estimated from the data itself, but treated as a known constant in the integral.

Since there are many factors that influence the prior distribution, such as the nature of the data, the parameterisation and the channel characteristics, the prior distribution on the mean is theoretically a sum of a large number of random variables. According to the central limit theorem, the mean of a large number of independent random variables, each with finite mean and variance, will be approximately normally distributed. The form of the prior distribution chosen in this paper is hence the same as the model for the data, a multivariate normal distribution, as in [8].

In the case of a multivariate normal distribution being chosen as the model for the data as well as the prior, $p(\boldsymbol{X}|\boldsymbol{\lambda}, M)$ and $p(\boldsymbol{\lambda}|M)$ in the marginal probability integral given in (5) can be expressed as

$$p(\boldsymbol{X}|\boldsymbol{\lambda}, M) = \prod_{i=1}^{N} \frac{|\boldsymbol{r}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{m})^T \boldsymbol{r} (\boldsymbol{x}_i - \boldsymbol{m})\right) \quad (8)$$

and

$$p(\boldsymbol{\lambda}|M) = \frac{|\boldsymbol{\tau}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{\mu})^T \boldsymbol{\tau} (\boldsymbol{m} - \boldsymbol{\mu})\right) \quad (9)$$

respectively, where $\boldsymbol{m}$ and $\boldsymbol{r}$ are the mean vector and precision matrix of the data, $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ are the mean vector and precision matrix of the prior and D is the dimensionality of the feature vector. Considering only the mean vector $\boldsymbol{m}$ as the variable of integration, (5) becomes

$$p(\boldsymbol{X}|M) = \int \prod_{i=1}^{N} \left[\frac{|\boldsymbol{r}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{m})^T \boldsymbol{r} (\boldsymbol{x}_i - \boldsymbol{m})\right) \right] \cdot \frac{|\boldsymbol{\tau}|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{\mu})^T \boldsymbol{\tau} (\boldsymbol{m} - \boldsymbol{\mu})\right) d\boldsymbol{m} \,. \quad (10)$$

While there is currently no known closed form solution to the indefinite integral given in (10), the definite integral over the entire space (ie. from $-\infty$ to $+\infty$) is known and can be derived with the assistance of an appropriate table of integrals, such as [9].

For the ease of evaluating the integral, given that $\boldsymbol{r}$ and $\boldsymbol{\tau}$ are full precision matrices, simultaneous diagonalization was used to transform the feature vector space so that $\boldsymbol{\tau}$ is whitened and $\boldsymbol{r}$ is diagonalized simultaneously in this new space. Each dimension can then be treated independently when evaluating the integral. Simultaneous diagonalization is achieved by finding a transformation matrix $\boldsymbol{A}$, to transform the data such that $\boldsymbol{X}' = \boldsymbol{A}\boldsymbol{X}$, where $(')$ denotes that the variable is expressed in the new space. As a result of this data transformation, $\boldsymbol{A}^T \boldsymbol{\tau} \boldsymbol{A} = \boldsymbol{I}$ is the identity matrix and $\boldsymbol{A}^T \boldsymbol{r} \boldsymbol{A} = \boldsymbol{r}'$ is diagonal. The expression for the exact value of the integral is given in (11), where $\boldsymbol{\Lambda}$ is the vector of eiganvalues of the prior covariance matrix in the original space. Note that the whitening of $\boldsymbol{\tau}$ and the diagonalization of $\boldsymbol{r}$, as opposed to the contrary, is a deliberate choice. Although both approaches will result in exactly the same marginal probability, whitening $\boldsymbol{\tau}$ allows the whitening matrix to be calculated only once. Conversely, whitening $\boldsymbol{r}$ would result in the need to calculate the whitening matrix for every segment being clustered. Whitening $\boldsymbol{\tau}$ is hence less computationally expensive, and the time required for clustering using this approach is comparable to the BIC clustering approach used in the baseline system.

3.4. Estimating the Hyperparameters

This paper proposes that the prior mean and precision can be estimated from the data itself. The prior mean estimate is given by the sample mean of all speech regions within the audio for a given show, $\boldsymbol{\mu} = \overline{\boldsymbol{X}}$, and the prior precision estimate for each segment is given by the sample precision matrix calculated from all speech regions of the audio scaled by the number of samples in the segment, $\boldsymbol{\tau} = N\boldsymbol{\Sigma}_X^{-1}$, using the central limit theorem as a guide. It is hypothesized in this paper that estimating the prior from the data itself will ensure a true and accurate representation of the data distribution, and alleviate the problem present in the BIC approximation, where the inferred prior may not accurately reflect the nature of the data being clustered.

4. RESULTS

This section presents the results of the Bayes Factor based clustering approach, as obtained on the RT-02 Evaluation dataset, and compares the results to the baseline system. The RT-02 Evaluation dataset consists of 6 recorded broadcast news shows, each with a scorable region of approximately 600 seconds.

4.1. Experimental Setup

To evaluate the proposed clustering strategy, the baseline system, using BIC clustering as the first clustering stage, is compared to a similar system with the BIC clustering stage replaced by Bayes Factor based clustering. The intermediate results obtained at the end of the first clustering stage are compared directly, as well as the resultant diarization performance of the respective systems as a whole.

For the direct comparison between intermediate clustering results at the end of the first clustering stage, average frame-level cluster purity and cluster coverage, as defined in [1], are calculated across all clusters. For a given cluster, cluster purity is defined as the proportion of speech that belongs to the dominant speaker over the total number of frames in the cluster. Cluster coverage takes into account the dispersion of a speaker's speech across clusters. For a given speaker, it is defined as the proportion of frames in the cluster which has the most amount of speech belonging to the speaker, over the total number of speech frames belonging to the speaker.

The effect of Bayes Factor based clustering on a diarization system as a whole can be evaluated using the Diarization Error Rate (DER) measure, as defined in [5]. The DER is the primary performance evaluation metric used in the NIST RT Diarization tasks. It can be interpreted as the percentage of the total amount of scorable time that is not attributed to the correct speaker, taking into account speech detection errors.

4.2. Clustering Results

Table 1 below shows the cluster purity and coverage results for each show at the end of the first clustering stage, using BIC and Bayes Factor based clustering respectively. The initial segments were produced using the baseline system processed up until the BIC clustering stage. Before clustering, the cluster purity and coverage values are 97.0% and 45.5% respectively. The low

$$p(\boldsymbol{X}|M) = \prod_{d=1}^{D} \sqrt{\frac{(\boldsymbol{r}'_d)^N}{(2\pi\boldsymbol{\Lambda}_d)^N(N\boldsymbol{r}'_d+1)}} \exp\left[\frac{-\boldsymbol{r}'_d}{2(N\boldsymbol{r}'_d+1)}\left[\left(\sum_{i=1}^{N}(\boldsymbol{x}'_{id}-\boldsymbol{\mu}'_d)^2\right)+\boldsymbol{r}'_d\left(N\sum_{i=1}^{N}((\boldsymbol{x}'_{id})^2)-\left(\sum_{i=1}^{N}(\boldsymbol{x}'_{id})\right)^2\right)\right]\right] \quad (11)$$

cluster coverage value is expected here, since speakers' utterances are dispersed throughout the audio as speaker segments at this stage. The cluster purity would ideally be 100% at this stage, and the loss is due to a small number of missed boundaries in the speaker segmentation stage. The results presented in Table 1 is based on the operating points for each system; it corresponds to the optimal stopping threshold for this clustering stage, emperically tuned on each system to produce the best possible DER on this dataset.

The average result of the 6 shows shown in Table 1 is calculated based on a time weighted average of the amount of scorable time in each show. Examining the results of each show individually, it is evident that the Bayes Factor system is able to perform more merges than the BIC system, without reducing the cluster purity. This is a desirable attribute, as it suggests that the Bayes Factor system is able to cluster the segments further, thus bringing the system to a more complete clustering state, without introducing further clustering errors.

Table 1. Results of Individual Shows - BIC vs Bayes Factor

Show	Coverage (%)		Purity (%)		# Merges	
	BIC	BF	BIC	BF	BIC	BF
1	61.0	61.2	95.2	95.5	169	173
2	68.6	65.8	98.3	98.3	152	155
3	97.8	76.3	99.2	99.2	88	83
4	80.8	78.2	90.0	90.3	81	86
5	62.3	60.6	98.3	98.5	132	135
6	63.3	63.1	96.8	96.1	81	87
Avg/Total	**71.9**	**67.4**	**96.2**	**96.3**	**703**	**719**

4.3. Diarization Performance Results

The overall diarization performance of the two systems, evaluated using the DER measure, is shown in Table 2. The average DER reported is also time weighted according to the length of the scorable region within each individual show. Compared to the BIC system, the differences in the outcome of the first clustering stage brought by the Bayes Factor system resulted in a considerable improvement in the overall DER in shows 2, 5 and 6. However, in shows 1 and 4, the BIC system marginally outperformed the Bayes Factor system. In show 1, the differences in the output of the first clustering stage, due to the termination of the initial clustering at different stages of the clustering process, resulted in different clusters being passed into the second viterbi realignment stage. This produced minor discrepancies in the locations of segment boundaries between the two systems, resulting in slight variations in DER, despite the overall clustering result being identical at the end of the second clustering stage. Overall, a 14.7% relative improvement in DER was achieved.

The results presented above suggests that the cluster coverage measure at the end of the first clustering stage does not necessarily provide a useful indication of diarization performance. Despite a decreased cluster coverage, the Bayes Factor system achieved an improved DER compared to the baseline. This suggests that the second clustering stage is able to recover the lower coverage by completing the clustering process. On the other hand, the ability to perform more merges without decreasing the cluster purity appears to have a direct impact on diarization performance. This reflects the importance of having pure, homogeneous segments at the end of the first clustering stage, as low cluster purity cannot be recovered in the second clustering stage.

Table 2. Comparison of Overall Diarization Results

Show	BIC System	Bayes Factor System
1	11.01	11.13
2	15.78	8.61
3	0.35	0.35
4	15.85	16.93
5	5.50	4.59
6	18.87	15.89
Avg DER	**11.46**	**9.77**

5. CONCLUSION

This paper proposes the use of the Bayes Factor to replace the BIC as a criterion for speaker clustering within a speaker diarization system. Since the BIC is only an approximation to the marginal likelihood of the data given the model, this paper proposes that the performance of a speaker clustering system can be enhanced by evaluating the Bayes Factor directly, to improve the accuracy and consistency by avoiding the error introduced in the BIC approximation. Results obtained on the RT-02 Evaluation dataset shows an improved clustering performance using Bayes Factor clustering, leading to an improvement in the overall Diarization performance.

6. ACKNOWLEDGEMENTS

This research was supported by an Australian Research Council (ARC) Linkage Grant No: LP0991238.

7. REFERENCES

[1] C. Barras, Z. Xuan, S. Meignier, and J. Gauvian, "Multistage speaker diarization of broadcast news," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 1505–1512, 2006.

[2] J. Gauvian, L. Lamel, Y. Kercadio, and G. Adda, "Transcription and indexation of broadcast data," *International Conference of Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, pp. 1663–1666, 2000.

[3] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," *Broadcast News Transcription and Understanding Workshop*, pp. 127–132, Feb 1998.

[4] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 1557–1565, 2006.

[5] J. Fiscus, "Fall 2004 rich transcription (rt-04f) evaluation plan," *National Institute of Standards and Technology*, 2004.

[6] C. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, 2006.

[7] D. Weakliem, "A critique of the bayesian information criterion for model selection," *Socialogical Methods and Research*, vol. 27, pp. 359–397, 1999.

[8] B. Ripley, *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[9] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed. San Diego: Academic Press, 2000.